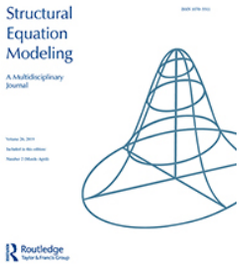
Routledge
Taylor & Francis Group



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

# Where Have the Persons Gone? – An Illustration of Individual Score Methods in Autoregressive Panel Models

Katinka Hardt, Martin Hecht, Johan H. L. Oud & Manuel C. Voelkle

**To cite this article:** Katinka Hardt, Martin Hecht, Johan H. L. Oud & Manuel C. Voelkle (2019) Where Have the Persons Gone? – An Illustration of Individual Score Methods in Autoregressive Panel Models, Structural Equation Modeling: A Multidisciplinary Journal, 26:2, 310-323, DOI: 10.1080/10705511.2018.1517355

**To link to this article:** https://doi.org/10.1080/10705511.2018.1517355

*Structural Equation Modeling: A Multidisciplinary Journal*, 26: 310–323, 2019
ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2018.1517355



# Where Have the Persons Gone? – An Illustration of Individual Score Methods in Autoregressive Panel Models

Katinka Hardt,[1] Martin Hecht,[1] Johan H. L. Oud,[2] and Manuel C. Voelkle[1,3]

[1]*Humboldt-Universität zu Berlin*
[2]*Radboud University Nijmegen*
[3]*Max Planck Institute for Human Development*

Much effort has been made to develop models for longitudinal data analysis, but comparably less attention has been paid to the use of individual specific values on latent variables in longitudinal models. In a tutorial style, this article introduces the reader to four common approaches to obtain individual scores – individual mean score, Bartlett method, regression method, Kalman filter – and reviews criteria commonly used to evaluate their performance. By means of simulated data, we mimic realistic scenarios and investigate in how far analytic results on the asymptotic performance of individual scores translate into practical situations. We end this article with a discussion of the use and usefulness of individual scores.

**Keywords**: longitudinal autoregressive models, individual diagnostics, individual scores (factor scores, sum score), Kalman filter

Where have the persons gone in longitudinal research? Many times, they disappear by being subsumed under averages or coefficients of variation. In psychology, averages or coefficients of variation are used to describe empirical data and to test theories about populations. But what if we are interested in the development of one particular individual from such a population? Then we need individual scores. The idea of an individual score is to map an individual onto a latent random variable.[1] In psychological research, constructs derived from and well-founded in psychological theory are often conceived of as latent variables. For longitudinal research, latent variables are often used to represent the "error free" construct at a given point in time (e.g., in autoregressive models) or may represent model parameters such as random intercepts or random slopes (e.g., in mixed/multilevel or latent growth curve models). In the present paper, we will focus on the former and will limit ourselves to normally distributed latent variables. Main purposes of using individual scores include individual diagnosis, monitoring or prediction.

To illustrate the purpose of individual scores, let us consider an example: imagine that there is a school district, which has implemented its own monitoring system into the schools in order to track the students' negative emotionality as this is known to affect school outcomes and dropout (e.g., Valiente, Swanson, & Eisenberg, 2012). Such a monitoring may aim at optimally fostering the students based on an individualized education program and to provide interventions if necessary. With latent growth curve or autoregressive models, we could analyze the development of negative emotionality at the group level of the students. But what if we are interested in the trajectory of one particular student of this cohort, say, of

Correspondence should be addressed to Katinka Hardt, Humboldt-Universität zu Berlin, Berlin, Germany katinka.hardt@hu-berlin.de



Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/HSEM.

Supplemental materials for this article can be found online at www.tandfonline.com/HSEM.

[1]For different conceptualizations of latent variables in psychology, see Bollen (2002).

Benni? To answer this question, a broad range of different approaches to scoring individuals is available, which differ in their statistical properties, and their choice should be carefully considered.

## PURPOSES OF THIS ARTICLE

The purposes of this article are twofold. First, we offer an introduction to the topic of individual scores. This might, for instance, be interesting for applied researchers and novices to this topic. We aim at raising the awareness of issues that should be considered when deciding on the use of individual scores and we want to put researchers in a position to make an informed decision about the use and usefulness of individual scores. The need to do so was already recognized decades ago by Horn (1965), and by doing so, we will extend existing similar endeavors (e.g., DiStefano, Zhu, & Mindrila, 2009; Glass & Maguire, 1966) by adding a longitudinal perspective. Second, we would like to stimulate further research in this field by pointing to promising procedures that are less commonly used in psychological research. This might be interesting to researchers who are already familiar with individual score approaches or who conduct research in this area.

The structure of this article is as follows: we first introduce the reader to the idea of individual scores and present selected methods to compute them. Next, we provide an overview of common criteria to evaluate individual score method performance and summarize findings on the performance of the selected methods in a longitudinal context. We then transfer existing analytical results on the performance of individual score methods to practical situations by means of an illustration based on simulated data. We end by discussing the use and usefulness of individual scores in particular situations.

One of our primary motives for writing this article relies on the observation that the sum score (sometimes also called "total score" or "composite score") continues being used, very often in inappropriate situations, either as mere sum score or as a variant of it, called item parcels (see Little, Cunningham, Shahar, & Widaman, 2002; Meade & Kroustalis, 2006; Nasser & Wisenbaker, 2003; Rhemtulla, 2016). Using sum scores, however, may be problematic because inherent to their use are three sources of potentially severe biases (e.g., Lastovicka & Thamodaran, 1991): first, it is implicitly assumed that the items measure a single construct (i.e., unidimensionality), second, it is ignored that one item may be more closely related to the underlying construct than another, and third, the amount of error due to unreliable measurement and the true score part that both compose the observed score cannot be disentangled. Thus, it is unknown how much we can "trust" the sum score. As approaches that make those assumptions more explicit and that account for the shortcomings of the sum score, we present the regression method (Thomson, 1938; Thurstone, 1934), the Bartlett method[2] (Bartlett, 1937), and the Kalman filter (Kalman, 1960). In contrast to the regression and the Bartlett method, the Kalman filter is less well known in psychological research. It is adopted from the field of engineering, where one also encounters the 'problem' of optimizing individual scores (see Priestley & Subba Rao, 1975 for the connection between Kalman filtering and factor analysis). Originally, it was developed to improve the tracking of systems (e.g., rockets in space) while integrating new data (e.g., on the rocket's actual position, velocity or direction). For this reason, the Kalman filter is called an "online" estimator. Along with the Kalman filter comes the Kalman smoother, which 'smoothes' back in time. In contrast to the Kalman filter, the Kalman smoother also uses information from future time points to optimally estimate an individual score at earlier time points. As the Kalman smoother is usually employed when the data collection is finished, it is called an "offline" estimator. For instance, if there are two measurement occasions, data from the second time point are used to estimate an individual score at the first time point. Thus, every individual score estimate at previous time points may change to the degree that data from a new measurement become available. In earlier research (e.g., Dolan & Molenaar, 1991; Oud, Jansen, van Leeuwe, Aarnoutse, & Voeten, 1999), it has been shown that the Kalman smoother can be formulated as a special case of the regression method. For this reason, we do not consider the Kalman smoother in the following. With this mixture of different approaches, we seek to open the reader's awareness about alternatives and about the statistical properties they have, while acknowledging that different methods may be differently useful in a given context.

## WHAT IS AN INDIVIDUAL SCORE? – A FORMAL ANSWER

For the purpose of the present paper, we define an individual score as a realization of a normally distributed random latent variable that conceptually represents a psychological construct. So, let any construct (e.g., any competencies, depression, positive/negative affect, etc.) be measured by $i = 1, \ldots, I$ multiple indicators (synonym: items), for which we can observe responses $y_i$. Further, let $c = 1, \ldots, C$ be the latent variables representing the constructs of interest. Each of the $j = 1, \ldots, J$ individuals has values on these latent variables, so-called *true scores*, that are contained in the vector $\mathbf{f}_j$. As $\mathbf{f}_j$ cannot be directly observed, *individual scores* $\hat{\mathbf{f}}_j$ need to be obtained by means of a statistical method.

There are several ways to map the observable responses $\mathbf{y}_j$ onto an individual's score $\hat{\mathbf{f}}_j$ and there exists a

[2] The regression method is also referred to as empirical Bayes estimate and individual scores based on the Bartlett method are maximum likelihood estimates (see Skrondal & Rabe-Hesketh, 2004).

controversy (e.g., Bartholomew, 1987) whether to conceive of this "act" of assigning values to latent variables as "prediction of a random variable" or as "estimation of a realized value of a random variable" (Robinson, 1991, p. 28). We will not take a stance at this discussion. In line with most of the research conducted in this field, we will use "estimation". This choice also facilitates to distinguish the *estimation* of individual scores (i.e., the act of assigning values on latent variables to persons) from the *prediction* of individual scores in the future. From this perspective, individual scores are not an immediate result of model estimation but require a second step after model estimation.

## HOW CAN WE OBTAIN INDIVIDUAL SCORES?

In general, approaches differ with regard to the type of information they incorporate when computing the individual scores and how this information is used. In this section, we present four common approaches to obtain individual scores and their standard errors: the sum score, the regression method, the Bartlett method, and the Kalman filter.

The simplest approach to obtain individual scores $\hat{\mathbf{f}}_j$ is to compute the *sum score* over the $I$ indicators in order to obtain $C$ individual scores:

$$\underset{C\times 1}{\hat{\mathbf{f}}_{\mathrm{SS}_j}} = \underset{C\times I}{\mathbf{S}'} \times \underset{I\times 1}{\mathbf{y}_j}, \tag{1}$$

where $\hat{\mathbf{f}}_{\mathrm{SS}_j}$ denotes individual scores $\hat{\mathbf{f}}_j$ obtained by computing individual sum scores, and $\mathbf{S}$ is a selection matrix that assigns a particular element in $\mathbf{y}_j$ to the corresponding construct it is supposed to measure. Readers who are less familiar with matrix notation might better recognize the computation of the sum score for a unidimensional construct (i.e., $C = 1$) from $\hat{f}_{\mathrm{SS}_j} = \sum_{i=1}^{I} y_{ij}$, where $y_{ij}$ is the response of individual $j$ to item $i$. Note that as long as there are no missing values, an individual's sum score and mean score as computed by $\hat{f}_{\mathrm{MS}_j} = \frac{\hat{f}_{\mathrm{SS}_j}}{I}$ for a unidimensional construct are perfectly correlated. Without relying on latent variable model parameters, we can calculate the standard error for the sum score according to $\frac{s}{\sqrt{I}}$, where $s$ is the standard deviation of the sum scores in a sample. This is in fact the standard error of a test's mean but adapted to the individual. The basic "ingredient" (or unit) of this test's mean are individual sum scores (or mean scores, respectively) in classical test theory; taking this standard error just transfers the idea of capturing random fluctuations for the mean to individual sum scores while assuming a constant standard error across individuals. Alternatively, the standard error of measurement according to $s \cdot \sqrt{1 - Rel(y)}$ could be used, where $Rel(y)$ is the reliability of this test; however, this strongly depends on the method itself to compute reliability.

Unlike the sum score, the regression method, the Bartlett method as well as the Kalman filter are based upon parameters estimated within a latent variable framework. When conceiving of our construct of interest as a latent variable (e.g., a factor), observable responses $\mathbf{y}_j$ to the manifest items are commonly linked to the underlying latent factor $\mathbf{f}_j$ by

$$\underset{I\times 1}{\mathbf{y}_j} = \underset{I\times 1}{\boldsymbol{\nu}} + \underset{I\times C}{\boldsymbol{\Lambda}} \times \underset{C\times 1}{\mathbf{f}_j} + \underset{I\times 1}{\boldsymbol{\varepsilon}_j}, \tag{2}$$

where $\mathbf{y}_j$ is a vector of manifest variables for person $j$, $\boldsymbol{\nu}$ contains the intercepts, $\boldsymbol{\Lambda}$ is the loading matrix connecting manifest and latent variables $\mathbf{f}_j$ and errors $\boldsymbol{\varepsilon}_j$ in the measurement model, with $\boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \Theta)$. If, in addition, relations among the latent variables are postulated, those can be expressed by

$$\underset{C\times 1}{\mathbf{f}_j} = \underset{C\times 1}{\boldsymbol{\alpha}} + \underset{C\times C}{\mathbf{A}} \times \underset{C\times 1}{\mathbf{f}_j} + \underset{C\times 1}{\boldsymbol{\zeta}_j}, \tag{3}$$

where $\boldsymbol{\alpha}$ contains the intercepts, $\mathbf{A}$ contains the structural coefficients and $\boldsymbol{\zeta}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. By having $\mathbf{f}_j$ both on the left handside and on the right handside of the equation, Equation (3) thus relates the latent variables in a model with each other.

Further, we need to consider that the variance-covariance matrix of the latent variables may not be equal to the variance-covariance matrix of individual scores, that is, $\mathbb{E}\left[\left(\hat{\mathbf{f}} - \mathbf{f}\right)\left(\hat{\mathbf{f}} - \mathbf{f}\right)'\right] > 0$ (e.g., Skrondal & Laake, 2001). Usually, this difference is referred to as "estimation" or "prediction" error of individual scores. In order to be in line with the term "individual score estimates" as used before, we will consequently use "individual score estimation error" here. For approaches, which incorporate a latent variable model (i.e., the regression and the Bartlett method as well as the Kalman filter), the principle of standard error calculation for individual scores is to take the square root of the diagonal elements of the estimation error variance-covariance matrix $\mathbf{P} = \mathbb{E}\left[\left(\hat{\mathbf{f}} - \mathbf{f}\right)\left(\hat{\mathbf{f}} - \mathbf{f}\right)'\right]$ (e.g., Oud, van den Bercken, & Essers, 1990), with $\hat{\mathbf{f}}$ representing the individual scores obtained by a particular method and $\mathbf{f}$ representing the true scores of the latent variables.

As an estimate of $\mathbf{f}$ the *regression method* (Thomson, 1938; Thurstone, 1934) provides us with

$$\underset{C\times 1}{\hat{\mathbf{f}}_{\mathrm{R}_j}} = \underset{C\times C}{\boldsymbol{\Phi}} \times \underset{C\times I}{\boldsymbol{\Lambda}'} \times \underset{I\times I}{\boldsymbol{\Sigma}^{-1}} \times \underset{I\times 1}{\mathbf{y}_j} \tag{4}$$

as individual score with $\boldsymbol{\Sigma}$ being the variance-covariance matrix of the observed variables and $\boldsymbol{\Phi}$ being the variance-covariance matrix of the latent variables $\mathbf{f}$. As a longitudinal model 'connects' the latent variables over time, the $\boldsymbol{\Phi}$ matrix allows us to incorporate longitudinal information. For the regression method,

$$\underset{C\times C}{\mathbf{P}_{\mathrm{R}}} = \underset{C\times C}{\boldsymbol{\Phi}} \times \left( \underset{C\times C}{\mathbf{I}} + \left( \underset{C\times I}{\boldsymbol{\Lambda}'} \times \underset{I\times I}{\boldsymbol{\Theta}^{-1}} \times \underset{I\times C}{\boldsymbol{\Lambda}} \right) \times \underset{C\times C}{\boldsymbol{\Phi}} \right)^{-1} \quad (5)$$

(Lawley & Maxwell, 1971, p. 109). The square root of the diagonal elements are used as standard errors for the individual scores. Two things become obvious here: first, relying on matrix algebra facilitates to recognize which model parameters are used to weigh the observed responses in $\mathbf{y}_j$ and, thus, to compute individual score estimates. Second, we notice that we can compute individual score estimates as soon as measurement and structural model parameter estimates are available. For this, it does not matter how those were estimated, for instance, either within the structural equation modeling (SEM) framework or within the state space modeling framework (see Chow, Ho, Hamaker, & Dolan, 2010, as well as Hunter, 2017, for a comparison of the two).

The *Bartlett method* (Bartlett, 1937) is defined by

$$\underset{C\times 1}{\hat{\mathbf{f}}_{\mathrm{B}_j}} = \left( \underset{C\times I}{\boldsymbol{\Lambda}'} \times \underset{I\times I}{\boldsymbol{\Theta}^{-1}} \times \underset{I\times C}{\boldsymbol{\Lambda}} \right)^{-1} \times \underset{C\times I}{\boldsymbol{\Lambda}'} \times \underset{I\times I}{\boldsymbol{\Theta}^{-1}} \times \underset{I\times 1}{\mathbf{y}_j} \; . \quad (6)$$

In contrast to the regression method, only the measurement model components $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ enter the computation. The $\boldsymbol{\Phi}$ matrix, whose elements include the structural information, is not part of the equation. For the Bartlett method

$$\underset{C\times C}{\mathbf{P}_{\mathrm{B}}} = \left( \underset{C\times I}{\boldsymbol{\Lambda}'} \times \underset{I\times I}{\boldsymbol{\Theta}^{-1}} \times \underset{I\times C}{\boldsymbol{\Lambda}} \right)^{-1} \quad (7)$$

(Lawley & Maxwell, 1971, p. 110), where the square root of the diagonal elements are used as standard errors for the individual scores.

An inherently longitudinal approach is the *Kalman filter* (Kalman, 1960). Its main idea is to improve estimation by integrating new incoming information in addition to the prediction based on past information. Transferring this idea to our initial example of tracking Benni's negative emotionality, we could first make model-based predictions and then compare this prediction with the actual measurement of his negative emotionality as soon as we collect the new data (for an application of the Kalman filter to students' development of decoding speed, see Oud et al., 1999). Thus, the principle of the Kalman filter is to optimally combine a model-based prediction with the arrival of new data from measurement in two steps. In the first step (prediction step), the individual score $\hat{\mathbf{f}}_{\mathrm{KF}}$ at a time point $t$ is predicted by the individual score at the previous time point $t-1$ yielding

$$\underset{C\times 1}{\hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t-1}}} = \underset{C\times 1}{\boldsymbol{\alpha}} + \underset{C\times C}{\mathbf{A}} \times \underset{C\times 1}{\hat{\mathbf{f}}_{\mathrm{KF}_{j,t-1|t-1}}} \, , \quad (8)$$

where $\boldsymbol{\alpha}$ contains intercepts and $\mathbf{A}$ denotes the transition matrix, which connects $\hat{\mathbf{f}}$ over time. $\mathbf{A}$ reflects the strength of the relationship between adjacent measurement occasions: the closer the absolute values in $\mathbf{A}$ are to one, the stronger the relationship and the better the prediction of $\hat{\mathbf{f}}$ at time point $t$ by $\hat{\mathbf{f}}$ at $t-1$.[3] This part corresponds to the structural part of a model in the SEM framework. The amount of uncertainty inherent in the prediction step is

$$\underset{C\times C}{\mathbf{P}_{\mathrm{KF}_{t|t-1}}} = \underset{C\times C}{\mathbf{A}} \times \underset{C\times C}{\mathbf{P}_{\mathrm{KF}_{t-1|t-1}}} \times \underset{C\times C}{\mathbf{A}'} + \underset{C\times C}{\boldsymbol{\Psi}_{t-1|t-1}} \; . \quad (9)$$

Note that the index for the time point in the Kalman filtering approach goes from $t=2$ to $T$, where $T$ denotes the total number of measurement occasions. As the Kalman filter is a recursive procedure, it requires to specify initial values for $\hat{\mathbf{f}}_{\mathrm{KF}_{1|1}}$ and $\mathbf{P}_{\mathrm{KF}_{1|1}}$. We used $t=1$ to denote the initial time point and thereby refer to the first measurement occasion. Basically, a researcher can either arbitrarily set values for $\hat{\mathbf{f}}_{\mathrm{KF}_{1|1}}$ and $\mathbf{P}_{\mathrm{KF}_{1|1}}$ to initialize the Kalman filter or make an "educated guess", for instance by inserting corresponding estimates from alternative individual score approaches (for an in-depth study of different initialization conditions, see Losardo, 2012). The choice of initialization was analytically shown to have an impact on the statistical properties of the Kalman filter results (Oud et al., 1999).

With the arrival of data from the new measurement at time point $t$ the prediction from time point $t-1$ is updated (update step) according to

$$\underset{C\times 1}{\hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t}}} = \underset{C\times 1}{\hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t-1}}} + \underset{C\times I}{\mathbf{K}_{t|t}} \times \left( \underset{I\times 1}{\mathbf{y}_{jt}} - \underset{I\times 1}{\hat{\mathbf{y}}_{j,t|t-1}} \right), \quad (10)$$

with $\hat{\mathbf{y}}_{j,t|t-1}$ being the responses predicted by $\boldsymbol{\Lambda} \times \hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t-1}} + \mathbf{v}$.

For the variance-covariance matrix of the Kalman estimation errors we get

$$\underset{C\times C}{\mathbf{P}_{\mathrm{KF}_{t|t}}} = \left( \underset{C\times C}{\mathbf{I}} - \underset{C\times I}{\mathbf{K}_{t|t}} \times \underset{I\times C}{\boldsymbol{\Lambda}} \right) \times \underset{C\times C}{\mathbf{P}_{\mathrm{KF}_{t|t-1}}} \quad (11)$$

with Kalman gain

[3] Note that the Kalman filter is part of the model parameter estimation in the state space modeling framework. Chow et al. (2010) as well as Oud et al. (1990) show that the state space framework is closely related to the structural equation modeling framework if a longitudinal model with autoregressive structure is postulated. In this paper, we restrict the Kalman filtering approach as used here to the estimation of individual scores, but it can also be used to estimate model parameters.

$$\underset{C\times I}{\mathbf{K}_{t|t}} = \underset{C\times C}{\mathbf{P}_{\mathrm{KF}_{t|t-1}}} \times \underset{C\times I}{\boldsymbol{\Lambda}'} \times \underset{I\times C}{(\boldsymbol{\Lambda}} \times \underset{C\times C}{\mathbf{P}_{\mathrm{KF}_{t|t-1}}} \times \underset{C\times I}{\boldsymbol{\Lambda}'} + \underset{I\times I}{\boldsymbol{\Theta}'})^{-1} \tag{12}$$

The Kalman gain determines how strongly the new measurement is weighted as compared to the prediction based on the previous time point. If the new measurement is rather unreliable, that is, has large error variances in $\boldsymbol{\Theta}$ in the measurement model, the update of $\hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t}}$ and $\mathbf{P}_{\mathrm{KF}_{t|t}}$ is more strongly driven by the prediction based on the previous time point. In contrast, if the new measurement is reliable with small error variances in $\boldsymbol{\Theta}$, the measurement's contribution to the updates of $\hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t}}$ and $\mathbf{P}_{\mathrm{KF}_{t|t}}$ is more strongly weighted.

Regarding the updated individual score $\hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t}}$, the difference $\mathbf{y}_{jt} - (\boldsymbol{\Lambda} \times \hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t-1}} + \boldsymbol{\nu})$ reflects a comparison of the actual measurement $\mathbf{y}_{jt}$ with the model-based predicted measurement $\boldsymbol{\Lambda} \times \hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t-1}} + \boldsymbol{\nu}$. Weighted with the Kalman gain (i.e., the reliability of the measurement), this difference adjusts ("updates") the individual score $\hat{\mathbf{f}}_{\mathrm{KF}_{j,t|t-1}}$ predicted from the previous time point.

Regarding the updated variance-covariance matrix $\mathbf{P}_{\mathrm{KF}_{t|t}}$, the term $\mathbf{I} - \mathbf{K}_{t|t} \times \boldsymbol{\Lambda}$ decreases as the reliability of the new measurement increases. As a consequence, the impact of the model-based predicted variance-covariance matrix $\mathbf{P}_{\mathrm{KF}_{t|t-1}}$ on $\mathbf{P}_{\mathrm{KF}_{t|t}}$ is downweighted the more reliable the new measurement is. The square root of the diagonal elements of $\mathbf{P}_{\mathrm{KF}_{t|t}}$ can then be used as standard errors to construct confidence intervals around the Kalman filter based individual scores.

## HOW CAN WE CHOOSE AN INDIVIDUAL SCORE METHOD?

After having introduced different methods for individual score estimation, we now turn to the question of how to choose between them. In the following, we present criteria that were used in prior research to evaluate the performance of the individual score methods.

*Strength of the relationship between the estimated individual score $\hat{\mathbf{f}}_j$ and the true score $\mathbf{f}_j$.* In previous research, this criterion takes different forms. In its most general form, it is defined as $\mathbb{E}(\hat{\mathbf{f}}\mathbf{f}') = \mathbb{E}(\mathbf{f}\mathbf{f}') = \boldsymbol{\Phi}$ (see Oud et al., 1990). That is, the covariances between the individual scores and the true scores are assumed to reflect the covariances between the true scores. If latent variables are standardized and assumed to be uncorrelated, then $\mathbb{E}(\hat{\mathbf{f}}\mathbf{f}') = \mathbb{E}(\mathbf{f}\mathbf{f}') = \mathbf{I}$. This special case is also referred to as *univocality* (Grice, 2001; Grice & Harris, 1998). Standardizing all variables and considering only the diagonal elements of $\mathbb{E}(\hat{\mathbf{f}}\mathbf{f}')$ gives us the correlations $r_{\hat{f}f}$. These correlations are sometimes presented as an index of *reliability* (e.g., Estabrook & Neale, 2013), sometimes as an index of *validity* (Grice, 2001; Heise & Bohrnstedt, 1970; Susmilch & Johnson, 1975). This index describes how well the relative positioning of individuals based on their true scores is maintained by individual score estimates. It is important to note that in latent variable modeling, individual scores are per se indeterminate as the number of unknown pieces of information due to unique (the residuals $\boldsymbol{\varepsilon}$) and common latent variables ($\mathbf{f}$) exceeds the number of available pieces of information (i.e., the observed items). This problem is usually referred to as *indeterminacy* of individual scores, and different indices of indeterminacy exist (e.g., Acito & Anderson, 1986). In simulation studies, $(r_{f\hat{f}_R})^2$ provides information on the "actual squared multiple correlation" (Acito & Anderson, 1986, p. 115). For a researcher who wants to compare individuals with each other based on their scores, an individual score method which has high correlations between individual score estimates and true scores appears suitable. What this criterion does not take into account is the absolute correspondence of each $\hat{f}_j$ with its true value $f_j$, which is accomplished by the next criterion: the bias.

*Bias*. The *bias* is the expected difference between an individual score estimate and the true score, $\mathbb{E}(\hat{f}_j) - f_j$. In our simulation, it is calculated for each person $j$'s individual score as $\frac{1}{N_\mathrm{r}}\sum_{r=1}^{N_\mathrm{r}}(\hat{f}_{j_r} - f_j)$, where $r = 1, \ldots, N_\mathrm{r}$ denotes the number of replications in the simulation. This property of an individual score method is particularly relevant if individual scores are used for diagnostic purposes. This might be the case, for instance, if there is an absolute threshold value, which determines whether a subject is eligible for additional support measures. When the goal is to make a decision about any given individual, methods that yield small bias are desirable.

*Variance*. The variance of an individual score estimate is defined as $VAR(\hat{f}_j) = E\left[\left(\hat{f}_j - E(\hat{f}_j)\right)\left(\hat{f}_j - E(\hat{f}_j)\right)'\right]$. It indicates the variability of an estimator due to random sampling error. In a simulation study, the variance can be captured by the variance or standard deviation of a particular score $\hat{f}_j$ over the number of replications: $VAR(\hat{f}_j) = \frac{1}{N_\mathrm{r}-1}\sum_{r=1}^{N_\mathrm{r}}(\hat{f}_{j_r} - \overline{\hat{f}_j})^2$ or as $SD(\hat{f}_j) = \sqrt{VAR(\hat{f}_j)}$, respectively. If we further assume that $\mathbb{E}(\hat{f}_j) = \overline{\hat{f}_j} = f_j$, we can also calculate the variance as $VAR(\hat{f}_j) = \frac{1}{N_\mathrm{r}}\sum_{r=1}^{N_\mathrm{r}}(\hat{f}_{j_r} - \overline{\hat{f}_j})^2$. The *precision* of an estimate is simply a transformation of the variance according to

$prec(\hat{f}_j) = \frac{1}{VAR(\hat{f}_j)}$. In general, individual score methods that yield very precise scores are desirable. That means that if we were able to repeat the analysis for the same subjects, we would obtain nearly identical individual scores based on a method with high precision.

*Mean squared error* (*MSE*). This criterion equals the averaged squared difference between a subject $j$'s individual score and the corresponding true score, $MSE(\hat{f}_j) = \mathbb{E}\left[(\hat{f}_j - f_j)^2\right]$. In a simulation study, the *MSE* can be calculated by $\frac{1}{N_r - 1}\sum_{r=1}^{N_r}(\hat{f}_{jr} - f_j)^2$. It is also common to take the square root of the *MSE*, denoted as *root mean squared error* (*RMSE*) with $RMSE = \sqrt{MSE}$, which combines bias and precision. If there is no bias, the *MSE* is equal to the variance of the individual score of subject $j$. In practice, there is usually a trade-off between bias and precision (see Geman, Bienenstock, & Doursat, 1992 for an illustration of the bias/variance dilemma). In a recent study, Curran, Cole, Bauer, Hussong, and Gottfredson (2016) used the *RMSE* to investigate individual score method quality in models that include background variables.

*Structure preservation*. Finally, another common criterion is the degree to which individual score methods are *structure preserving*, sometimes also called *correlation preserving* in case of standardized variables (e.g., Oud et al., 1990; Saris, de Pijper, & Mulder, 1978); that is, the degree to which they satisfy the constraint $\mathbb{E}(\hat{\mathbf{f}}\,\hat{\mathbf{f}}') = \mathbb{E}(\mathbf{f}\mathbf{f}') = \mathbf{\Phi}$. A special case often considered for standardized variables in exploratory factor analysis (EFA) is $\mathbb{E}(\hat{\mathbf{f}}\,\hat{\mathbf{f}}') = \mathbb{E}(\mathbf{f}\mathbf{f}') = \mathbf{I}$, which is then called *orthogonality*. The criterion of structure preservation is often indirectly applied, for instance, when the focus is on the use of individual scores to further investigate structural relationships between latent variables (e.g., Devlieger, Mayer, & Rosseel, 2016; Skrondal & Laake, 2001). In this case, statistical properties of structural coefficients (e.g., regression coefficients) obtained from analyses based on individual scores are of focal interest, for instance, the degree to which $\mathbf{f}^* = \hat{\mathbf{A}}\mathbf{f}^* + \boldsymbol{\zeta}$ is equivalent to $\mathbf{f} = \mathbf{A}\mathbf{f} + \boldsymbol{\zeta}$. Here, $\mathbf{f}^*$ denotes the estimate $\hat{\mathbf{f}}$ as obtained in a first step. In a second step, structural coefficients in $\hat{\mathbf{A}}$ are estimated based on $\mathbf{f}^*$, where $\mathbf{f}^*$ is taken as fixed; $\boldsymbol{\zeta}$ contains the error terms. In this line of research, the question is how well $\hat{\mathbf{A}}$ matches $\mathbf{A}$ across different individual score methods. It is obvious that for this purpose the structure preserving property in the first step is of prime importance.

Depending on the framework of analysis (e.g., EFA vs. theoretically derived models with structural components accounting for the longitudinal structure), these criteria are differently important. Traditionally, most of them were applied within an EFA framework. In contrast, longitudinal modeling requires imposing constraints on the measurement models to ensure measurement invariance as well as making explicit assumptions of how different measurements are related over time (e.g., linearly or quadratic, etc.). As a consequence, the mapping of manifest onto latent variables (via measurement models) and the relationship among the latent variables (via the structural model) need to be explicitly specified. Therefore, criteria to be considered in a longitudinal context primarily include bias, precision, *MSE*, and $r_{\hat{f}f}$ rather than structure preservation or univocality. Next, we will turn to what we know from previous research about the use of individual scores in longitudinal contexts.

## WHAT DO WE KNOW ABOUT THE USE AND PERFORMANCE OF INDIVIDUAL SCORE METHODS IN PSYCHOLOGICAL LONGITUDINAL STUDIES?

Only little is known about the performance of individual scores in longitudinal psychological studies. Seminal works for the $J = 1$ case were conducted for instance by Molenaar (1985), or, for the panel case with $J > 1$, by Oud et al. (1990, 1999). Besides a comprehensive introduction to Kalman filtering, the authors of the latter studies analytically derive properties of individual scores based on the regression method, the Bartlett method, and the Kalman filter. Building upon Lawley and Maxwell (1971), they show that the regression method yields biased scores but with minimum variance, whereas the Bartlett method leads to unbiased individual scores but with a larger variance. The Kalman filter is considered an optimal estimator, but its performance depends on the initialization method. In case of an initialization method that is optimal in terms of unbiasedness *and* minimum variance, the Kalman filter would yield optimal scores without any restriction. In practice, however, such a generally optimal initialization does not exist. Accordingly, properties of the Kalman filter scores first depend on the initialization until they "forget" about the initialization when time passes and new data arrive (e.g., Visser & Molenaar, 1988). Oud et al. (1999) showed that the Kalman filter has minimum variance already at $t = 2$, even though the Kalman filter was initialized with the Bartlett estimator at $t = 1$ which does not have minimum variance. This is true for most situations except in the case that the Bartlett estimator is based on extremely large error variances in the measurement model. Having discussed the asymptotic performance of the Kalman filter, the regression and the Bartlett method, we now turn to the question of how analytically derived

asymptotic properties translate into finite sample scenarios.

## ILLUSTRATION USING SIMULATED DATA

We illustrate the properties of the previously presented individual score methods by means of simulated data for our fictitious example of Benni as introduced at the beginning. We consciously decided against an empirical example as the computation of individual scores itself can be easily done as soon as we estimated a model. Simulated data allow us to measure the extent to which individual scores obtained by different methods match the true values. Although asymptotic performance has been derived analytically as discussed before, Monte Carlo simulations open the door to study realistic scenarios and to investigate practical implications. For instance, we learned from analytical derivations that individual scores based on the regression method do have minimum variance but are biased, that the Bartlett method yields individual scores that have minimum variance among all approaches which are unbiased (Lawley & Maxwell, 1971), and that the Kalman filter forgets about its initialization and quickly becomes the approach with minimum variance. But what does this mean for finite samples and, in particular, for specific individuals within a sample and their individual trajectories? How well does an individual score obtained by a particular method match the underlying true score when there are differently reliable measurements or differently persistent processes as it is often the case in typical psychological applications?

## PROCEDURE

The simulation had three steps: data generation, model estimation, and computation of individual scores. These three steps were replicated $N_r = 500$ times per condition. Data generation and the estimated model were identical in all conditions. The rationale behind this procedure is that plugging model parameter estimates into formulas for individual scores relies on the assumption that the estimated parameters are valid for the whole sample. In practice, we would first postulate a model based on substantive theory, then estimate the model parameters and finally interpret the model results only if the postulated model sufficiently fits the empirical data. In case of a valid model we can proceed to compute individual scores in a very last step, based on the estimated model parameters for each data set according to the methods presented before. Thereby, the computation of individual scores is independent of the model parameter estimation. We summarized these measures over replications and per condition and analyzed them with respect to the outcome criteria described later. All analyses were conducted using the software R (R Core Team, 2017). Model estimation was carried out using the package OpenMx (Boker et al., 2011; Neale et al., 2016), individual scores were computed with our own routines. In the online supplemental material, we provide example code for data generation according to the model described next, for model estimation, and for obtaining individual scores using lavaan (Rosseel, 2012), OpenMx (Boker et al., 2011; Neale et al., 2016), and Mplus (Muthén & Muthén, 2017).

## MODEL

Back to our hypothetical scenario, we assume that Benni is part of a cohort that comprises $J = 200$ elementary school students whose negative emotionality is assessed on $T = 5$ time points during a school year with equidistant time intervals. We account for the longitudinal structure by using an autoregressive model of order one, AR(1), which connects negative emotionality at adjacent time points by

$$f_{jt} = a \cdot f_{j,t-1} + \zeta_{jt} \tag{13}$$

with $\zeta_{jt}$ having variance $\Psi_t$. We assume that the relationships between adjacent measurements are constant over time by imposing equality constraints on the autoregression coefficients such that $\mathrm{a}[t, t-1] = \mathrm{a}$, and we further assume that the process is stationary (for the concept of stationarity, see e.g., Hamilton, 1994, pp. 45–46). Equation (13) thus describes the generation of a subject $j$'s "true trajectory" (based on "true scores" at every $t$).

Further, in our hypothetical scenario, negative emotionality is a latent construct that is measured by $I = 5$ manifest indicators. For this reason, a measurement model according to Equation (2) is additionally incorporated in the data generation. Assuming that negative emotionality is measured with the same instrument (i.e., the same manifest indicators) at every measurement occasion, we constrain the loadings as well as the error variances in the measurement models to be equal over time for each indicator (e.g., the entries $\lambda_{it}$ in the $\mathbf{\Lambda}$ matrix reduce to $\lambda_i$, and $VAR(\varepsilon_{it})$ in $\mathbf{\Theta}$ are constrained to $VAR(\varepsilon_i)$). We thus establish measurement invariance over time and ensure that the latent variable is on the same scale at different time points. All variables are standardized and the initial variance of $f$ is constrained to one for reasons of scaling and model identification. Figure 1 depicts a conceptual path diagram of the model that was used for data generation and model estimation.

## DESIGN

We experimentally varied model parameters in both the measurement part ($\lambda_i$ and $VAR(\varepsilon_i)$) and in the structural part of the autoregressive model (autoregression coefficient $a$) to study their effects on individual scores. In the measurement

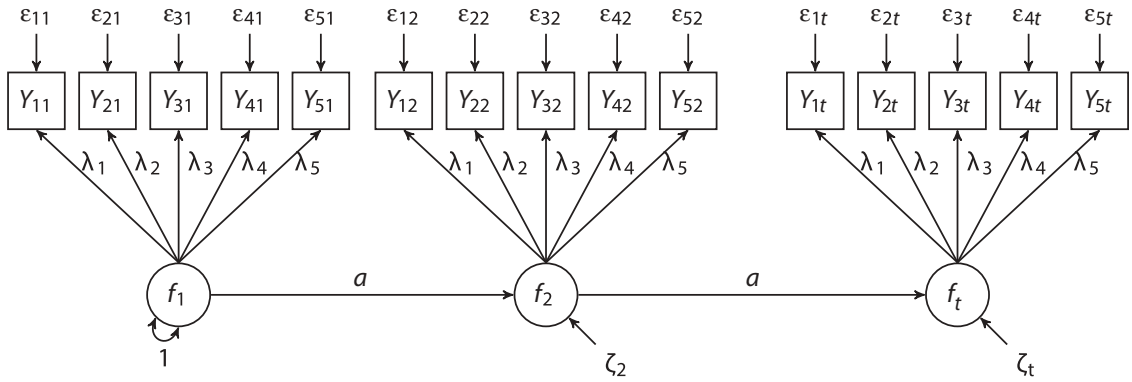



FIGURE 1 Conceptual path diagram of an autoregressive model of order one, five observed indicators and measurement invariance assumed.

part, we chose the loadings for our five observable indicators in such a way that their mean was either 0.6 ("reliability 1" condition) or 0.8 ("reliability 2" condition). As standardized items have a variance of 1, a loading of 0.6 corresponds to 64% ($1 - 0.6^2 = 0.64$) residual variance in this item and a loading of 0.8 corresponds to 36% residual variance in this item. Further, we varied the variation in the loadings (0 vs. 0.13). In conditions with a variance of 0.13, loadings were chosen in such a way that they were symmetric around the mean. We expect that variation in loadings has an impact on the performance of the individual mean score as this method equally weighs all items and therefore ignores any variance in the loadings. In the structural part of the model, the magnitude of the autoregression coefficient $a$ was either 0.25 or 0.75 representing different degrees of persistence. With regard to our initial example, a low $a$ coefficient indicates lack of persistence in the sense that negative emotionality at one time point is not predictive for negative emotionality at a subsequent time point. In contrast, a high $a$ coefficient indicates persistence, that is, negative emotionality at one time point is highly predictive for negative emotionality at the next time point. Table 1 summarizes the conditions.

With this illustrative simulation, our goal is to assess how individual score methods behave when we pretend that we could take one and the same sample of subjects and test it repeatedly under different conditions. What varies then is the error in the measurement model, not the true trajectories of the individuals, except when we change structural components. That is, we first draw a value (true score) from a standard normal distribution for each individual. Then, for a given autoregressive coefficient $a$, we compute their true trajectory. We only do so once for each $a$-condition and keep the true trajectories constant over all the other conditions and over replications given a particular $a$. We thus repeatedly 'expose' our sample to differently reliable measurements by manipulating parameters in the measurement model (see Table 1), and we do so for two sets of conditions resulting from two differently persistent processes ($a = 0.25$ and $a = 0.75$). As the generation of true trajectories only requires a structural model but no measurement model, it

TABLE 1
Overview of Conditions: Variation in Loadings for Each Set of Persistence Conditions ($a = 0.25$ and $a = 0.75$)

| *Condition* | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ | $M_\lambda$ | $SD_\lambda$ |
|---|---|---|---|---|---|---|---|
| *Reliability 1:* | | | | | | | |
| M06,SD000 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.00 |
| M06,SD013 | 0.60 | 0.45 | 0.55 | 0.65 | 0.75 | 0.60 | 0.13 |
| *Reliability 2:* | | | | | | | |
| M08,SD000 | 0.80 | 0.80 | 0.80 | 0.80 | 0.80 | 0.80 | 0.00 |
| M08,SD013 | 0.80 | 0.65 | 0.75 | 0.85 | 0.95 | 0.80 | 0.13 |

*Note*. Condition labels are concatenated means and standard deviations of the loadings.

becomes obvious that we obtain two sets of true scores for the 200 subjects, one set for the $a = 0.25$ conditions and one set for the $a = 0.75$ conditions. In total, we thus have $2 \cdot 200 = 400$ unique individuals. Using the same seed, we obtain the same true scores at $t = 1$ for the two $a$-conditions. Because of the different autoregression coefficients $a$, we obtain different true scores for $t > 1$. Therefore, Benni, who only is Benni as defined by his unique individual trajectory, can only be Benni in repeated testing under differently reliable measurement conditions, but not in the presence of a differently persistent negative emotionality.

## COMPUTATION OF INDIVIDUAL SCORES

We computed individual scores according to four of the previously presented approaches (abbreviations as used in the presentation of the results are given in parentheses): the regression method (Regression), the Bartlett method (Bartlett), the individual mean score (MeanScore)[4] and three versions of the

[4] Note that we use the individual mean score here rather than an individual sum score in order not to change the metric of the responses that was used in the data generation. Otherwise, this metric would be altered by each item that additionally enters the sum score, thus, sum score

Kalman filter. As the Kalman filter is a recursive procedure, it requires initialization. We included three different ways of initialization: first, random initialization by drawing values from a uniform distribution characterized by the minimum and by the maximum of the observed data (KF). The variance of the Kalman estimation error was set to 0.5 for initialization. We chose this large value to assess the performance of the Kalman filter under particularly unfavorable conditions. Second, we used individual scores and estimation error obtained from the regression method for the first time point to initialize the Kalman filter (KFiniR). Third, we used individual scores and estimation error obtained with the Bartlett method to initialize the Kalman filter (KFiniB). The computation of individual scores according to all methods but the mean score required the incorporation of model parameters as estimated beforehand. We constructed 95% confidence intervals around the individual scores using the standard errors as described above.

## OUTCOME CRITERIA

For our study, we rely on criteria that provide information of individual score method performance on the subject level as opposed to aggregate performance indices such as *structure preservation*. To start analyses on the individual level, we first inspect *correlations between a particular individual score method and the corresponding true scores*. For each replication, we first calculated Pearson's product-moment correlation coefficient $r$, then Fisher-$Z$-transformed these correlations, averaged them, and transformed them back into $r$ in order to facilitate interpretation. This criterion provides information on the relative ordering of the individuals. A value of 1 indicates a perfect correlation, that is, the positioning of an individual relative to another one is perfectly maintained by the estimate. In a next step, we shift the focus from the relative positioning of subjects to the absolute match of a score obtained by different methods with the true score. In order to scrutinize in how far the deviation of an individual score from the underlying true score is relevant in practical situations, we additionally incorporated a criterion based on 95% confidence intervals. For each replication, we first calculated dichotomous indicators which indicate whether the confidence interval of an individual score captured the true score (score 1) or not (score 0) per person, and per time point. We call this the *mismatch criterion*, with $d = \begin{cases} 1 \text{ "mismatch"} \\ 0 \text{ "match"} \end{cases}$. Then we calculated the percentage of "mismatches" over replications for each condition. This criterion can be considered as a "reversed coverage" as commonly used in simulation studies. It indicates in how many replications a parameter estimate is not included in the corresponding confidence interval. The mismatch criterion borrows from real data situations by constructing a confidence interval around the individual scores per replication according to the formulas presented before. Then, the percentage of how many times over replications this confidence interval did not capture the true score is computed. We introduced the proposed terminology to facilitate a substantive interpretation in line with the focus of this article. The mismatch criterion thus reflects the amount of practical "mismatch" or "misclassification" that could occur in the simulated scenarios beyond the expected error probability of 5%. However, it is obvious that a large individual score estimation error, and, thus, wide confidence intervals, can mask bias; one should be aware of this interrelationship. Nevertheless, applying the mismatch criterion comes for the benefit of practical relevance and practical meaningfulness. Unlike all the other criteria we presented before, this outcome relates to practical situations as closely as possible. In contrast, the *MSE*, for instance, combines the criteria of bias and variance, but we do not know the threshold value of when an individual score method with a particular *MSE* leads to a different individual decision in practical settings. This is only accomplished by the mismatch criterion as it incorporates the individual score estimation error, which is the basis for obtaining standard errors to construct confidence intervals around the point estimate of an individual score. For this reason, we will first take a closer look at the standard errors of the individual score methods to build a sense of the precision of our individual score methods. To illustrate the mismatch criterion, Figure 2 shows the trajectories of different individual score methods for two exemplary individuals, Benni and Jacky. If Benni comes from a population whose negative emotionality is rather non-persistent ($a = 0.25$ conditions), the subfigure on the left describes his trajectory while being exposed to a measurement of a given quality. On the right, the trajectory of Jacky is shown. Jacky stems from a different population, a population whose negative emotionality is described by a persistent process with $a = 0.75$. Jacky has the same true score at $t = 1$ but different true scores at each $t > 1$. This figure makes two of our outcome criteria evident: the width of the confidence intervals around the individual scores as determined by the standard errors and whether individual score plus confidence interval capture the true score or not.

## RESULTS

The performance of individual score methods is evaluated in terms of their capability to maintain the relative ordering of individuals (correlation criterion) and in terms of whether the individual score confidence interval captures the true score or not (mismatch criterion). As the findings are very similar across time for each individual score method, we only report the results for $t = 4$ for the correlation criterion (Table 2) and for the standard errors (Table 3), which are related to the mismatch criterion. Figures showing the full results can be found in the

inherent bias and bias due to a different metric would be confounded. As noted before, the mean perfectly correlates with the sum score when $I$ is constant for all individuals as it is the case in our illustrative simulation.

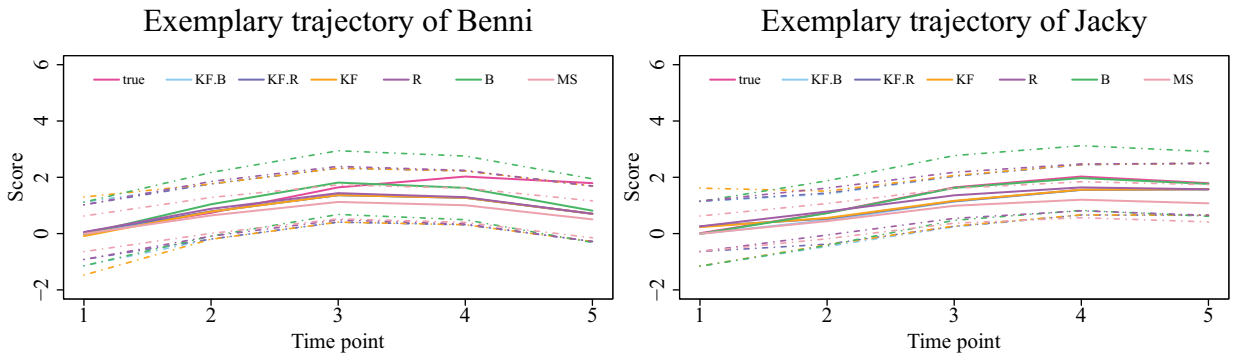



FIGURE 2 True = true score, KF = Kalman filter, KF.B = KFiniB, KF.R = KFiniR, R = Regression, B = Bartlett, MS = MeanScore. Exemplary trajectories of the true score and estimated scores by method for the M06,SD000 condition, for Benni (left panel) and Jacky (right panel).

TABLE 2
Correlation between Individual Scores and True Scores at $t = 4$ by Individual Score Method and Condition

| *Condition* | *Regression* | *Bartlett* | *MeanScore* | *KF* | *KFiniR* | *KFiniB* |
|---|---|---|---|---|---|---|
| $a = 0.25$ | | | | | | |
| M06,SD000 | 0.864 | 0.858 | 0.858 | 0.862 | 0.862 | 0.862 |
| M06,SD013 | 0.878 | 0.873 | 0.859 | 0.876 | 0.876 | 0.876 |
| M08,SD000 | 0.949 | 0.948 | 0.948 | 0.949 | 0.949 | 0.949 |
| M08,SD013 | 0.970 | 0.969 | 0.949 | 0.969 | 0.969 | 0.969 |
| $a = 0.75$ | | | | | | |
| M06,SD000 | 0.914 | 0.873 | 0.873 | 0.897 | 0.899 | 0.899 |
| M06,SD013 | 0.922 | 0.887 | 0.874 | 0.908 | 0.909 | 0.909 |
| M08,SD000 | 0.963 | 0.954 | 0.954 | 0.959 | 0.959 | 0.959 |
| M08,SD013 | 0.977 | 0.973 | 0.955 | 0.975 | 0.975 | 0.975 |

*Note*. 'M06' in the condition labels refers to reliability 1 conditions, 'M08' refers to reliability 2 conditions. The correlation is close to zero for the randomly initialized Kalman filter in every condition.

TABLE 3
Mean Individual Score Standard Errors at $t = 4$ by Individual Score Method and Condition

| *Condition* | *Regression* | *Bartlett* | *MeanScore* | *KF* | *KFiniR* | *KFiniB* |
|---|---|---|---|---|---|---|
| $a = 0.25$ | | | | | | |
| M06,SD000 | 0.485 | 0.577 | 0.311 | 0.488 | 0.488 | 0.488 |
| M06,SD013 | 0.457 | 0.532 | 0.310 | 0.460 | 0.460 | 0.460 |
| M08,SD000 | 0.304 | 0.324 | 0.374 | 0.305 | 0.305 | 0.305 |
| M08,SD013 | 0.234 | 0.242 | 0.374 | 0.234 | 0.234 | 0.234 |
| $a = 0.75$ | | | | | | |
| M06,SD000 | 0.425 | 0.586 | 0.326 | 0.457 | 0.456 | 0.456 |
| M06,SD013 | 0.403 | 0.540 | 0.326 | 0.432 | 0.431 | 0.432 |
| M08,SD000 | 0.280 | 0.326 | 0.398 | 0.292 | 0.292 | 0.292 |
| M08,SD013 | 0.222 | 0.244 | 0.397 | 0.228 | 0.228 | 0.228 |

*Note*. 'M06' in the condition labels refers to reliability 1 conditions, 'M08' refers to reliability 2 conditions. KF was initialized with an estimation error of 0.5. The table shows the mean standard error over replications per condition. The *SD* of this standard error over replications is very small (with a maximum of 0.026) and therefore negligible.

Online supplemental material. In those figures, it becomes obvious that the impact of a random initialization of the Kalman filter has died out at $t = 4$ (at the latest) in every condition. Further, at $t = 4$, the regression method can still exploit information from future time points. The correlation and the mismatch criterion indicate three main results: first, all methods perform much better in reliability 2 conditions than in reliability 1 conditions. Second, if the process is persistent (i.e., $a = 0.75$ conditions), the randomly initialized Kalman filter (KF) takes 1-2 time points more than in case of a non-persistent process until its initialization impact has died out. Third, variation in the loadings (as indicated by SD013-conditions) leads to better performance (as compared to SD000-conditions) for all methods except the mean score. This effect is due to the mean score ignoring any differences in the strength of the relationship between the items and the latent variable (i.e., the loadings). In contrast, all methods that account for the loading structure differently weigh the corresponding responses. Responses to items that have high loadings are more strongly weighted such that the corresponding individual score estimate is more strongly determined by these responses. Especially in the reliability 2 conditions there is at least one response to an item with small measurement error. This response almost matches the true score and is most strongly weighted by the corresponding loading. As a consequence, individual scores obtained by methods that account for loadings are hardly contaminated by responses to items with large measurement error because the correspondingly small loadings downweigh these responses. Closer inspection of the correlation criterion (see Table 2) reveals that in case of a non-persistent process (i.e., $a = 0.25$ conditions), the only difference we can observe is the lower performance of the individual mean score when there is variation in the loadings. This is due to the equal weighing of the responses. In this situation, the relative ordering of the individuals is almost equally well maintained independent of which individual score method we choose. In contrast, if the process is persistent (i.e., $a = 0.75$ conditions), we observe a better relative positioning of individuals for methods that incorporate longitudinal information (i.e., the regression method and the Kalman filter versions). Inspecting the mismatch criterion in more detail further reveals that the individual score interval based on the

Bartlett method most closely and consistently matches the expected 5% across all conditions (see Figure 3 and the full results table in the Online supplemental material). As this is the average across persons within one condition, the comparatively small standard deviation indicates that those 5% of mismatch are true for most subjects in the sample. On average, the regression method and the Kalman filter versions have slightly higher mismatch rates with larger standard deviations (except for the randomly initialized KF). This means that for many individuals the individual score intervals of those methods do not capture the true score and, thus, may for instance lead to wrong individual diagnostic decisions. Further, we observe that the mean score is very sensitive to reliability 1 conditions having mismatch rates far beyond the expected 5%, that is, when error in the measurement model is relatively large. In this situation, the mean score and its confidence interval may lead to wrong decisions. Thus, we have seen that the individual score interval according to the Bartlett method most accurately reflects the nominal confidence level (i.e., here 95%). To develop a better understanding of the mismatch criterion, it is also worthwhile to consider the standard errors (see Table 3). As a reminder, standard errors as used here are calculated based on the estimation error for the Bartlett and the regression method as well as for the Kalman filter versions. For individual mean scores, we use the standard error for a test's mean but adapted it to individuals as described before. As a first result we observe that the standard errors for the mean score are slightly smaller in reliability 1 conditions than in reliability 2 conditions. As this standard error is calculated based on the standard deviation of the mean scores, we see here that in the presence of relatively large amounts of measurement error (reliability 1 conditions), the mean score is not capable of discriminating well between persons. Further, for the remaining individual score methods, we see that individual score methods benefit from incorporating longitudinal information (Regression, KF, KFiniR and KFiniB) with regard to their standard errors if the process is persistent ($a = 0.75$). With regard to our hypothetical scenario, it does not make a huge difference which of the approaches we choose to describe Benni's trajectory as long as error in the measurement model is not too large. If, in contrast, error in the measurement model is rather large, we should be aware that we are more likely to have a mismatch on the individual level especially when using the mean score, but also when using the regression method or the Kalman filter versions as compared to using the Bartlett method. As we know from inspecting the standard errors, this comes at the price of larger confidence intervals, and, thus, more imprecise individual scores.

## ON THE USE AND USEFULNESS OF INDIVIDUAL SCORES

Individual scores are important whenever an individual and his or her trajectory are of interest. Apart from the practical example of Benni, a range of purposes is conceivable, including monitoring, diagnosis, and prognosis. For instance, individual scores can be used to track the development of abilities or skills that are important in education (e.g., decoding speed, reading comprehension, etc.) or to monitor the progression of mental illnesses such as depression in a clinical context. In the same settings, when diagnostically relevant thresholds are defined, individual scores allow us to determine whether and when a subject reaches, exceeds or falls below such a threshold or even when this will be the case in the future (i.e., for prognosis). Such thresholds may stem

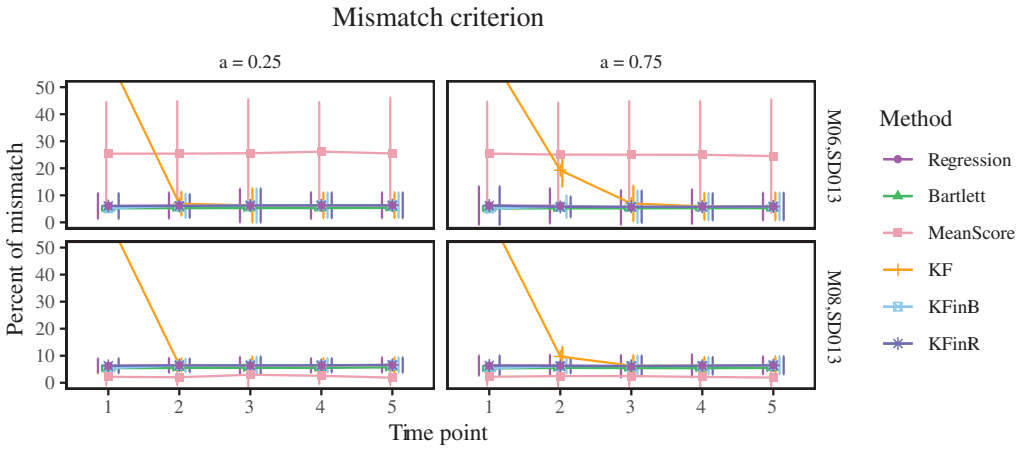



FIGURE 3 Percentage of mismatch (mean and *SD* across persons) of individual scores and true scores by time and individual score method (four conditions). For the KF, the percentage of mismatch at the first time point is between 60 and 63 (not displayed).

from comparing an individual's development to that of a reference group or by theoretical knowledge. In all of the simulated scenarios, there were no remarkable differences between the regression method, the Bartlett method and the Kalman filter versions (except for the impact of the random initialization of the KF). Even the individual mean score showed good performance, at least in high reliability conditions. However, the problem with the individual mean score is that it inherently relies on very strong, and oftentimes unrealistic assumptions regarding unidimensionality, loadings, and measurement error as explained above. In practice, those assumptions are hardly ever met. In real data situations, we never know the quality of our measurement unless we use latent variable models. We first need to check those assumptions before we can assess how much we can trust a particular individual score method. Our little illustrative simulation showed three main results: first, relatively large error in the measurement model generally leads to worse performance of individual score methods, second, methods that include model parameters (all methods but the mean score) benefit from variation in the loadings, and third, in the presence of relatively large measurement error, individual score methods perform worse when the process is non-persistent and when there is no variation in the loadings. Special caution should be exercised in reliability 1 conditions which exhibited the largest difference between methods.

In propensity score analysis, individual scores were used to statistically adjust for differences between control groups and intervention groups on fallible covariates (Raykov, 2012). However, the validity of this approach has been questioned (Lockwood & McCaffrey, 2016). In how far the different statistical properties of different individual score methods play a role in this context and in how far the consideration of individual score estimation error may add value to their usefulness in the balancing of groups remains an open question. Likewise, individual scores may be useful in latent interaction modeling (e.g., Kelava et al., 2011; Moosbrugger, Schermelleh-Engel, & Klein, 1997). Modeling the nonlinear multivariate distribution directly is considered the most recent, supreme approach in this context that outperforms traditional product indicator approaches (e.g., Kelava et al., 2011). Schumacker (2002) showed that individual scores are an easy-to-use alternative to product indicator approaches although the particular individual score method remains unmentioned as well as differential performance of different individual score methods. Whether and in which situations the use of individual scores can be considered a valuable alternative to situations in which distribution analytic approaches reach their limitations appears worthwhile to investigate (e.g., in small sample size situations).

When should we be cautious about the use of individual scores or not even resort to individual scores? When we are interested in structural parameters and when there is reason to doubt that the proposed structure we would like to model (e.g., regressive relationships between latent variables) validly describes our data (i.e., when there is need to test a model). One of the most valuable achievements of psychometric research during the past decades is latent variable modeling. Advances in latent variable modeling allow for a thorough investigation of the model-data fit, while at the same time being easy to use and available in nearly every mainstream software (e.g., Mplus, lavaan, OpenMx, Stata, SAS, etc.). When relying on individual scores to investigate structural relationships instead (very often also referred to as factor score regression), one runs into the danger of not discovering lack of model-data fit and of suggesting a valid model interpretation nonetheless. That said, in some selected and well-defined situations individual scores have proven to be a valuable alternative to modeling directly within a latent variable framework. To name a few examples, Hoshino and Bentler (2013) acknowledge the bias of regression coefficients based on individual scores and propose a correction method, with a particular focus on manifest indicators with different scales (e.g., categorical as well as continuous indicators). Building upon the method of Croon (2002), Devlieger et al. (2016) discuss a way to correct the bias of regression coefficients obtained from factor score regression. Devlieger and Rosseel (2017) extend this approach to path analysis. The main advantage over SEM approaches can be seen in the performance of this approach in small sample situations in terms of bias and convergence. However, this approach is not suited to test theoretically postulated structures among latent variables and to explain structures in empirical data.

The property of testing structures in data appears important for psychological research. How well a theoretical model fits empirical data is in the focus of residual analysis, another field of research in which individual scores have proven to be useful. An individual residual can be considered just as an individual: for each person, there are "true" unknown residuals that can be approximated by individual score methods. As many distributional assumptions of structural equation models concern the residuals, obtaining individual residuals allows for testing these assumptions and for conducting outlier and influential case analyses. In earlier studies, Bollen and Arminger (1991) demonstrated the use of the regression and the Bartlett method to conduct such analyses. More recently, Hildreth (2013) investigated both asymptotic and finite sample properties of residuals obtained by the regression, the Bartlett as well as the Anderson and Rubin method. In addition, Hildreth (2013) illustrates the use of residuals to identify outliers and influential observations in SEM building upon and extending the work by Bollen and Arminger (1991). Future work might focus on the sensitivity of different individual score methods to detect outliers and influential cases.

## ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the paper. Katinka Hardt is a pre-doctoral fellow of the International Max Planck Research School on the Life Course (LIFE, www.imprs-life.mpg.de; participating institutions: MPI for Human Development, Freie Universität Berlin, Humboldt-Universität zu Berlin, University of Michigan, University of Virginia, University of Zurich).

## FUNDING

We acknowledge support by the Open Access Publication Fund of Humboldt-Universität zu Berlin.

## REFERENCES

Acito, F., & Anderson, R. D. (1986). A simulation study of factor score indeterminacy. *Journal of Marketing Research*, *23*(2), 111–118. doi:10.2307/3151658

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London, UK: Charles Griffin; Oxford University Press.

Bartlett, M. S. (1937). THE STATISTICAL CONCEPTION OF MENTAL FACTORS. *British Journal of Psychology. General Section*, *28*(1), 97–104. doi:10.1111/j.2044-8295.1937.tb00863.x

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., & Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306–317. doi:10.1007/s11336-010-9200-6

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, *53*(1), 605–634. doi:10.1146/annurev.psych.53.100901.135239

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology*, *21*, 235. doi:10.2307/270937

Chow, S.-M., Ho, M. R., Hamaker, E. L., & Dolan, C. V. (2010). Equivalence and differences between structural equation modeling and state-space modeling techniques. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 303–332. doi:10.1080/10705511003661553

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195–223). Mahwah, NJ: Erlbaum.

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 827–844. doi:10.1080/10705511.2016.1220839

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: a comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770. doi:10.1177/0013164415607618

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM? *Methodology*, *13*(Supplement 1), 31–38. doi:10.1027/1614-2241/a000130

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11.

Dolan, C. V., & Molenaar, P. C. M. (1991). A note on the calculation of latent trajectories in the quasi Markov simplex model by means of the regression method and the discrete Kalman filter. *Kwantitatieve Methoden*, *38*, 29–44.

Estabrook, R., & Neale, M. (2013). A comparison of factor score estimation methods in the presence of missing data: Reliability and an application to nicotine dependence. *Multivariate Behavioral Research*, *48*(1), 1–27. doi:10.1080/00273171.2012.730072

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. doi:10.1162/neco.1992.4.1.1

Glass, G. V., & Maguire, T. O. (1966). Abuses of Factor Scores. *American Educational Research Journal*, *3*(4), 297–304. doi:10.3102/00028312003004297

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. doi:10.1037/1082-989X.6.4.430

Grice, J. W., & Harris, R. J. (1998). A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research*, *33*(2), 221–247. doi:10.1207/s15327906mbr3302_2

Hamilton, J. D. (1994). *Time series analysis*. Princeton, N. J: Princeton University Press.

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. *Sociological Methodology*, *2*, 104–129. doi:10.2307/270785

Hildreth, L. (2013). *Residual analysis for structural equation modeling*. Retrieved from http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=4407&context=etd

Horn, J. L. (1965). An empirical comparison of methods for estimating factor scores. *Educational and Psychological Measurement*, *25*(2), 313–322. doi:10.1177/001316446502500202

Hoshino, T., & Bentler, P. (2013). Bias in factor score regression and a simple solution. In A. De Leon & K. Chough (Eds.), *Analysis of Mixed Data* (pp. 43–61). Boca Raton, FL: Chapman and Hall/CRC.

Hunter, M. D. (2017). State space modeling in an open source, modular, structural equation modeling environment. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 1–18. doi:10.1080/10705511.2017.1369354

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45. doi:10.1115/1.3662552

Kelava, A., Werner, C. S., Schermelleh-Engel, K., Moosbrugger, H., Zapf, D., Ma, Y., & West, S. G. (2011). Advanced nonlinear latent variable modeling: distribution analytic lms and qml estimators of interaction and quadratic effects. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 465–491. doi:10.1080/10705511.2011.582408

Lastovicka, J. L., & Thamodaran, K. (1991). Common factor score estimates in multiple regression problems. *Journal of Marketing Research*, *28*(1), 105. doi:10.2307/3172730

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2 ed.). New York: American Elsevier Pub. Co.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 151–173. doi:10.1207/S15328007SEM0902_1

Lockwood, J. R., & McCaffrey, D. F. (2016). Matching and weighting with functions of error-prone covariates for causal inference. *Journal of the American Statistical Association*, *111*(516), 1831–1839. doi:10.1080/01621459.2015.1122601

Losardo, D. (2012). *An Examination of Initial Condition Specification in the Structural Equations Modeling Framework* (Unpublished doctoral dissertation). University of North Carolina at Chapel Hill.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*(3), 369–403. doi:10.1177/1094428105283384

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202. doi:10.1007/BF02294246

Moosbrugger, H., Schermelleh-Engel, K., & Klein, A. (1997). Methodological problems of estimating latent interaction effects. *Methods of Psychological Research*, *2*(2), 95–111.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, *63*(5), 729–757. doi:10.1177/0013164403258228

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535–549. doi:10.1007/s11336-014-9435-8

Oud, J. H. L., Jansen, R. A. R. G., van Leeuwe, J. F. J., Aarnoutse, C. A. J., & Voeten, M. J. M. (1999). Monitoring pupil development by means of the Kalman filter and smoother based upon SEM state space modeling. *Learning and Individual Differences*, *11*(2), 121–136. doi:10.1016/S1041-6080(00)80001-1

Oud, J. H. L., van den Bercken, J. H., & Essers, R. J. (1990). Longitudinal factor score estimation using the Kalman filter. *Applied Psychological Measurement*, *14*(4), 395–418. doi:10.1177/014662169001400406

Priestley, M. B., & Subba Rao, T. (1975). The estimation of factor scores and Kalman filtering for discrete parameter stationary processes. *International Journal of Control*, *21*(6), 971–975. doi:10.1080/00207177508922050

R Core Team. (2017). *R: A language and environment for statistical computing (version 3.4.0)*. Vienna, Austria: R Foundation for Statistical Computing.

Raykov, T. (2012). Propensity score analysis with fallible covariates: A note on a latent variable modeling approach. *Educational and Psychological Measurement*, *72*(5), 715–733. doi:10.1177/0013164412440999

Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods*, *21*(3), 348–368. doi:10.1037/met0000072

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, *6*(1), 15–32. doi:10.1214/ss/1177011926

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 2. doi:10.18637/jss.v048.i02

Saris, W. E., De Pijper, M., & Mulder, J. (1978). Optimal procedures for estimation of factor scores. *Sociological Methods & Research*, *7*(1), 85–106. doi:10.1177/004912417800700104

Schumacker, R. E. (2002). Latent variable interaction modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(1), 40–54. doi:10.1207/S15328007SEM0901_3

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. doi:10.1007/BF02296196

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Susmilch, C. E., & Johnson, W. T. (1975). Factor scores for constructing linear composites: Do different techniques make a difference? *Sociological Methods & Research*, *4*(2), 166–188. doi:10.1177/004912417500400202

Thomson, G. H. (1938). Methods of estimating mental factors. *Nature*, *141*(3562), 246. doi:10.1038/141246a0

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, *41*(1), 1–32. doi:10.1037/h0075959

Valiente, C., Swanson, J., & Eisenberg, N. (2012). Linking students' emotions and academic achievement: When and why emotions matter: emotion and achievement. *Child Development Perspectives*, *6*(2), 129–135. doi:10.1111/j.1750-8606.2011.00192.x

Visser, H., & Molenaar, J. (1988). Kalman filter analysis in dendroclimatology. *Biometrics*, *44*(4), 929–940. doi:10.2307/2531724